Statistical Potentials for RNA 3D Structures: From Base-Level to All-Atom Scoring

Negin Heidarifard

Master 2 GENIOMHE — Université Paris-Saclay

Abstract

We implemented a distance-based statistical potential to evaluate RNA 3D structures using experimentally solved conformations. Starting from a coarse-grained C3′–C3′ model, the approach was extended to an all-atom, base-only formulation. The resulting scoring engine is able to discriminate native RNA structures from predicted models, despite being trained on a very small dataset. This work illustrates how simple statistical principles already capture non-random geometric signatures of RNA folding.

1 Introduction

Predicting RNA three-dimensional structure remains a major challenge due to the enormous conformational space accessible to polynucleotide chains. Native RNA folds correspond to minima of the Gibbs free energy landscape, but computing this energy from first principles is not tractable in practice.

A classical alternative is to use *statistical potentials*, where effective interaction energies are learned from structural databases. The underlying assumption is that geometrical features observed frequently in native structures correspond to energetically favourable configurations.

In this project, we construct a statistical potential based on interatomic distance distributions in solved RNA structures. The model is first defined at the nucleotide level using C3′ atoms, and then extended to an all-atom, base-specific description.

2 Training Data

The potential was trained on three experimentally determined RNA structures downloaded from the Protein Data Bank:

- **1EHZ**: Hairpin ribozyme (native reference)
- **4TNA**: tRNA^{Asp}
- **6TNA**: tRNA^{Asp}

Although this dataset is very small, it is sufficient to highlight clear, non-random distance patterns in native RNA folds.

3 Methods

3.1 Distance Selection and Filtering

Only intrachain distances were considered. To avoid trivial local geometry imposed by covalent constraints, pairs of residues separated by fewer than three positions along the sequence were excluded.

Two levels of resolution were implemented:

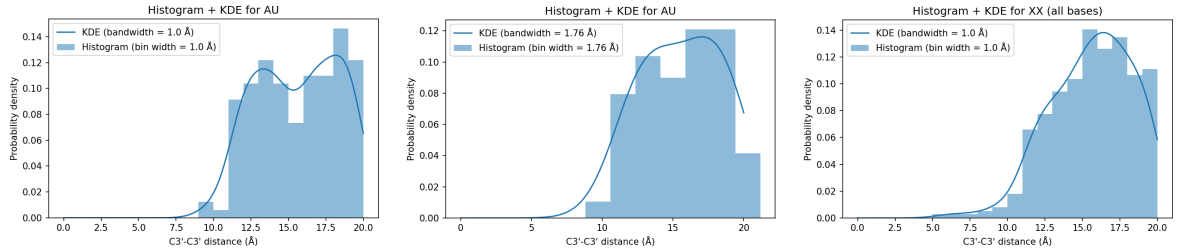
- **Base-level model**: C3′–C3′ distances between nucleotides.
- **All-atom model**: distances between all base atoms (38 atom types), excluding backbone atoms.

Distances larger than 20 Å were ignored.

3.2 Histogramming and KDE Analysis

Distances were discretised into bins of width $\Delta r = 1$ Å. To justify this choice, histograms were compared to Gaussian kernel density estimates (KDE).

Figure 1 shows that a bin width of 1 Å captures the main modes of the distribution while avoiding excessive noise. Increasing the bandwidth smooths the distribution but removes secondary features, supporting the choice of $\Delta r = 1$ Å for training.



(a) AU, $\Delta r = 1$ Å (b) AU, adaptive bandwidth (c) XX reference

Figure 1: Histogram and KDE analysis of C3′–C3′ distance distributions.

3.3 Statistical Potential

For each distance bin r, the pseudo-energy was defined as

$$u_{ij}(r) = -\log\left(\frac{f_{ij}^{\mathrm{OBS}}(r) + \varepsilon}{f_{XX}^{\mathrm{REF}}(r) + \varepsilon}\right),$$

where $f_{ij}^{\mathrm{OBS}}(r)$ is the observed frequency for a given pair type, and $f_{XX}^{\mathrm{REF}}(r)$ is the reference distribution pooling all pairs. A small pseudocount ε avoids division by zero, and values were clipped to a maximum of $+10$.

An alternative non-log (information-gain–like) formulation was also implemented and validated.

4 Scoring of RNA-Puzzles Models

The learned potentials were used to score RNA-Puzzles models by summing interpolated pseudo-energies over all valid pairs.

For the all-atom model, 741 distinct atom-pair types were learned. When scoring the RNA-Puzzles structure **1Y26**, the model considered **73 751** contributing atom pairs, yielding a total pseudo-energy of

$$\Delta G_{\text{all-atom}} = 13\,607.03 \quad \text{(arbitrary units)}.$$

This large number reflects the much finer resolution of the all-atom model compared to the C3′-only approach.

Structure	Role	# pairs	Pseudo-ΔG
1EHZ	Native	357	~ 4.4
1Y26	Puzzle	504	~ 135.5
2L8H	Puzzle	155	~ 99.2
5T5A	Puzzle	506	~ 231.5

Table 1: Scores obtained with the base-level statistical potential.

Native structures systematically receive much lower scores than predicted models, confirming the discriminative power of the potential.

5 Discussion

Despite its simplicity and limited training data, the statistical potential captures meaningful geometric signatures of RNA folding. The extension to an all-atom representation greatly increases sensitivity, at the cost of interpretability and scale.

Importantly, the model does not aim to reproduce physical Gibbs free energies, but rather to provide an effective objective function for ranking structures.

6 Limitations and Perspectives

The main limitations are the small training set, the absence of solvent and entropic effects, and the purely pairwise nature of the potential. Future extensions could include environment-specific potentials, stacking terms, or fully KDE-based continuous energy functions.

7 Conclusion

This project demonstrates that even minimal statistical models, when carefully constructed and validated, can effectively discriminate native RNA structures from non-native conformations. The final workflow constitutes a complete, reproducible RNA scoring engine from training to evaluation.